



Article

Learning Performance Prediction and Alert Method in Hybrid Learning

Huijuan Zhuang ¹ , Jing Dong ², Su Mu ^{3,*} and Haiming Liu ⁴ 

¹ International Business College, South China Normal University, Guangzhou 510631, China

² School of History and Culture, Qufu Normal University, Jining 273165, China

³ Institute of Artificial Intelligence in Education, South China Normal University, Guangzhou 510631, China

⁴ School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

* Correspondence: musu@m.scnu.edu.cn

Abstract: In online learning, students' learning data such as time and logs are commonly used to predict the student's learning performance. In a hybrid context, learning activities occur both online and offline. Thus, how to integrate online and offline learning data effectively for an accurate learning performance prediction becomes very challenging. This paper proposes a "prediction and alert" model for students' learning performance in a hybrid learning context. The model is developed and evaluated through analyzing the 16-week (one semester) attributes of English learning data of 50 students in the eighth grade. Six significant variables were determined as learning performance attributes, namely, qualified rate, excellent rate, scores, number of practice sessions, practice time, and completion. The proposed model was put into actual practice through four months of application and modification, in which a sample of 50 middle school students participated. The model shows the feasibility and effectiveness of data analysis for hybrid learning. It can support students' continuous online and offline learning more effectively.

Keywords: learning performance prediction; hybrid learning; ridge regression; clustering algorithm



Citation: Zhuang, H.; Dong, J.; Mu, S.; Liu, H. Learning Performance Prediction and Alert Method in Hybrid Learning. *Sustainability* **2022**, *14*, 14685. <https://doi.org/10.3390/su142214685>

Academic Editors: Gwo-Jen Hwang, Morris JONG Siu-yung, Haoran Xie and Gary Cheng

Received: 24 August 2022

Accepted: 2 November 2022

Published: 8 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hybrid online and offline learning has existed in China for over two decades, but has been limited in scale as merely trials or occasional experimentation. However, after the COVID-19 outbreak, hybrid learning has exploded in schools. It is mentioned in the "Evaluation" section of the US 2017 National Education Technology Plan "Reimagining the Role of Technology in Education: 2017 National Education Technology Plan Update" that, by improving the educational data system, educational leaders can use comprehensive data to improve the quality and efficiency of learning tools and resources supported by technology [1]. Applying AI technology to data analysis of hybrid learning can better support online and offline learning and build personalized digital learning experiences [2,3]. Teachers can use these data to provide effective information for teaching intervention and decision making, and to help all learners to obtain a personalized learning experience with better engagement and greater relevance. Environmental analysis based on learning performance to improve learning services for personalized learning is an important direction for educational development. The rapid emergence of learning analytics, educational big data, and artificial intelligence technologies that measure, collect, analyze, and report on the data generated by learners and their contexts offer technical possibilities for formative assessment and learning alerts.

As it integrates the merits of both learning modes, hybrid learning can be more flexible, diverse, and profound than either online or offline learning [4]. Hybrid learning is not a mere combination of face-to-face and online learning. It is a holistic learning process, which uses the best delivery method for the successful achievement of the learning objectives. The principle of hybrid learning is to form a unique learning experience on

the basis of combining the advantages of both, so as to adapt to the teaching situation and serve the teaching objectives [5]. To achieve an effective hybrid learning experience for both teacher and students can be very challenging because of the open, flexible, and independent learning process [6]. Take, for instance, the independent learning ability, which is the essential assurance for satisfactory learning outcomes in a hybrid learning context, when simultaneous supervision or guidance from the teacher is unavailable in the classroom. Given this situation, both teachers and students hope to obtain timely information about the learning performance based on the analysis of the learning process, behaviors, and achievements. Such information also facilitates customized instructions and precise teaching in a hybrid learning context. The ultimate goal of hybrid learning is to achieve the best learning effect and economic benefits.

Due to the difficulties in obtaining offline data, predicting the performance of hybrid learning becomes difficult. In addition, there are few valid indicators and algorithms for hybrid learning prediction, which diminishes the accuracy of learning performance predication [7]. It is detrimental to the teaching and learning practice if predictions and alerts are unreliable for the teachers and students. This research tries to improve the approach to learning performance prediction based on the state of the art. Using hybrid learning data, we created and evaluated a “Prediction-Alert” algorithm for hybrid learning performance. It is an important focus of educational development in the era of artificial intelligence to analyze students’ learning process data and improve learning services to promote personalized learning. The application of artificial intelligence technology and big data analysis technology in the field of education can promote precise teaching and adaptive learning, and effectively improve the process management of hybrid learning [8].

2. Related Work

As online teaching and learning are necessary during COVID-19 and retained post-pandemic for their perceived strengths, there have been some methods and algorithms for online learning prediction. This section reviews the existing research on learning performance prediction and early alerts. Then, as the main application of learning prediction, we introduce the line of research on ridge regression and the clustering algorithm.

2.1. Research on Students’ Learning Performance Prediction in Hybrid Environments

The task of predicting students’ performance includes giving records of a series of behaviors or activities that students have shown in the past and estimating their future state [9]. This is followed by dealing with the assessment of the student’s learning performance, providing course adaptation and learning recommendations based on the student’s learning behavior, approaches dealing with the evaluation of learning material and educational web-based courses, giving feedback to both teacher and students in e-learning courses, and developments for detection of atypical student learning behaviors [10].

On students’ learning performance prediction, related research has focused on the use of algorithms and learning data to generate quantitative values of learning metrics to predict learning states. Gong et al. [11], by analyzing students’ learning behaviors related to a certain course, based on the parallel computation and binary logistic regression algorithm in the Spark framework, created the off-line learning prediction model, through which, in the real-time environment, large-scale real-time learning prediction can be realized. Na [12] proposed a framework for predicting students’ learning performance based on a behavioral model and described their behavior characteristics and added context information to the collaborative filtering algorithm, including student knowledge point mastery and class knowledge points, and students’ mastery is predicted according to the learning path excavated. Lu, et al. [13] applied learning analytics and educational big data approaches, including proposed course, consisting of video viewing behaviors, out-of-class practice behaviors, homework and quiz scores, and after-school tutoring, for the early prediction of students’ final academic performance by principal component regression in a blended course. Moises et al. [14] used machine learning to create models for the early prediction

of students' performance in solving LMS assignments, by just analyzing the LMS log files generated up to the moment of prediction.

2.2. Research on Students' Learning Performance Early Alerts

Learning performance alerts can be summarized as a set of mechanisms, from data acquisition and performance prediction to early-alert display. It aims to make full use of learning data to build an effective model [15]. Hu et al. [16] proposed an alert system to help identify students at risk of dropping out or predict their learning achievements by analyzing the logs recorded in the Learning Management System (LMS). Through equal-width discretization, Gökhan et al. [17] used students' eBook reading data to develop an early alert system, using 13 prediction algorithms with the data from different weeks of the course to determine the best performing model and optimum time for possible interventions. Howard, et al. [18] presented findings from a statistics university course which has weekly continuous assessment and a large proportion of resources on the Learning Management System Blackboard, and identified weeks 5–6 (half way through the semester) as an optimal time to implement an early alert system, as it allows time for the students to make changes to their study patterns, whilst retaining reasonable prediction accuracy.

To sum up, studies on early learning alerts mainly focus on analyzing the interactive classification of students' learning process and students' learning support behavior through the log files of a learning management system, realizing multi-label learning and multi-task prediction of online learning relationships, exploring students' learning behavior performance by using various general intelligent technology algorithms, in addition to evaluating its impact on academic performance, as well as early prediction of academic performance by using learning analysis technology and learning data and behavior data generated by models, and analyzing the characteristic attributes related to students' performance at risk. However, reasonable hybrid learning early alert indicators and early alert algorithms have not yet been formed, which makes it difficult to guarantee the accuracy and rationality in the early alert process, and the early alert results produce large errors, which is not conducive to the teaching application of the analysis results. Therefore, in this study, the acquired data are first pre-processed, and alert models are constructed based on the analysis of the correlation between each attribute and the target variable, and early alert analysis is performed based on the prediction results. If the selected algorithm and tools can visualize the results, it enables the researcher to grasp the data more intuitively and accurately.

2.3. Research on Methods of Predicting Students' Learning Performance

Previously, in hybrid learning environments, linear models such as linear regression and logistic regression have always been the main research focus [19]. Some used more complex techniques such as decision tree, random forest, logistic regression, reverse neural network, clustering, and support vector machine, while other studies have already attempted to harness the predicted capabilities of neural networks [20]. Adoption, modification, and validation of warning models are important issues in learning performance prediction research. Take, for instance, Limsathit Wong et al. [21], who found the decision tree method more effective than a random forest in identifying dropouts when predicting the dropout rate of freshmen and sophomores from the Thailand National University of Technology. Xu et al. [22] used three machine learning algorithms—decision tree, neural network, and support vector machine—to predict academic performance by including online duration, Internet traffic volume, and connection frequency. Jiang et al. [20] categorized the learners according to the data of learning behaviors in six MOOCs offered by Peking University on Coursera. They further predicted the students' learning achievements by linear discriminant analysis, logistic regression, and linear kernel support vector machine. Matzavela et al. [23] created adaptive dynamic tests for assessing student academic performance and formulated a predictive model for students' knowledge level, according to

the weights of the decision tree. Table 1 summarizes the purpose and characteristics of the analytical methods used in previous studies.

Table 1. Purpose and Characteristics of Data Analysis Methods in Learning Prediction.

Prediction/Alert Method	Purpose	Characteristics
Linear regression	Learning achievement prediction	Free of complex calculations, fast modeling, and strong interpretability, but ill-fitting nonlinear data.
Logistic regression	Research on the influencing factors of learning achievements	Easy application and interpretation, applicable to continuous independent variables, but sensitive to the multicollinearity between independent variables.
Clustering algorithm	Research on modeling the influencing factors of learning achievements	Fast, simple, and flexible, but requiring a specified number of clusters; possible to generate low-quality clusters with inappropriate operation.
Random forest	Dropout rate prediction	Certain scalability and robustness when dealing with abnormal data, but sometimes unconstrained due to the tendency to over-fit single tree.
Bayesian network	Learning achievement prediction	Possible automatic expansion with the update of data sets, but sometimes replaceable due to the overly simple algorithm.
Time series algorithm	Learning achievement prediction	Very simple model, with only endogenous variables needed when alerting, but requiring highly stable time series data.
Artificial neural network	Learning achievement prediction	Highly intelligent algorithm, capable of finding the optimal solution at a high speed, but easy to lose information and requiring high data integrity.
Support Vector Machine	Learning achievement prediction	Modeling to the nonlinear decision boundary with nonlinear kernel function, but difficult to adjust parameters and with narrow applicability.

The algorithms listed in Table 1 are commonly used in the field of data mining and learning analysis to realize learning prediction. Among them, the linear regression algorithm is fast in modeling and has strong explanatory power, which is very effective in analyzing the relationship between small data quantities and is convenient in parameter adjustment. Based on the above characteristics of linear regression algorithms, this study uses a linear regression algorithm to build a learning performance prediction model of hybrid learners. At present, there are three methods to solve the multicollinearity problem in multiple linear regression models. First, in a step-by-step regression method, find out the independent variable that causes multicollinearity in the multiple regression model and exclude it. Second, the complex multiple linear regression model is divided into several simple univariate linear regression models, and the relationship between independent variables and dependent variables is explained from multiple dimensions. Third, the multicollinearity of multiple linear regression can be solved by the ridge regression method, which can reduce the variance of parameters. In this study, in order to keep the independent variables intact and ensure that the expressions of five independent variables to dependent variables can be obtained, the ridge regression algorithm is used to eliminate the multicollinearity of the multiple linear regression. Fast modeling and strong interpretability are features of the linear regression algorithm, which is used in data mining and data analysis for learning prediction. It is especially effective in analyzing the relationships between the data of small-sized samples, and fairly convenient in parameter adjustment. Given its characteristics, the study adopts a linear regression algorithm to construct the learning performance prediction model of hybrid learning.

3. Method and Experiments

Based on the above methods, in view of the characteristics of the hybrid learning process, the study aims to propose and validate a learning performance prediction and alert model of hybrid learning by integrating the data that reflect the learners' online, offline, and hybrid learning performances. The study is developed in two stages. In the first stage, the entire data set was used to determine the alert attributes of the learning performance, while the data of the testing group was used to construct the "Prediction-Alert" model. In the second stage, the data of the validating group were used to verify the effectiveness of the alert, and to modify the algorithm according to the analysis results.

The first stage used all 16 weeks of learning data to construct a "Predict-alert model". Questions need to be addressed. How do we determine the alert attributes for learning performance? How do we make a learning performance prediction model? Which classification algorithm can classify the values of learning performance attributes?

The second stage is concerned with the accuracy of the prediction and alert algorithm. The questions to be addressed include: To what extent can the algorithm accurately predict and alert to the learning performance based on the existing data? If the algorithm were not as accurate as expected, how could it be improved? (see Figure 1)

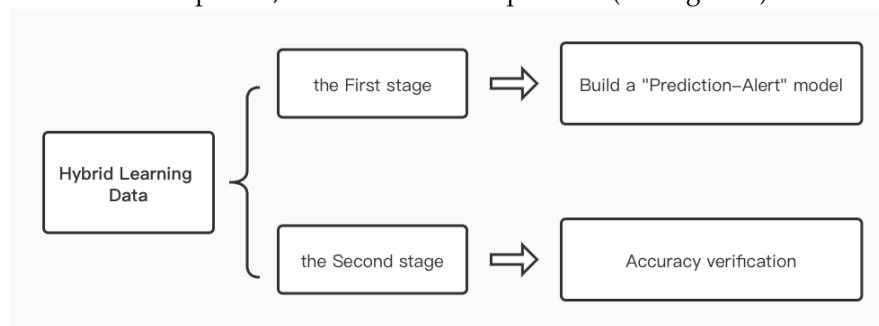


Figure 1. Overview of the research.

3.1. Data Collection

Online and offline hybrid learning is continuous. It is more important to predict the performance of students based on all the data of their learning process; so far, for each student, all data in the whole learning process were analyzed without time sampling. The data source for the study was the English learning records of 50 eighth graders on a hybrid learning platform. These records reflect students' participation in regular English language learning, such as listening, speaking, reading, and writing in English, either online or in the classroom. Every observation unit in the whole population has the same opportunity to be selected into the sample and has the same opportunity to be grouped. The purpose of randomization is to make the experimental group and the control group comparable by randomizing and balancing the influence of interference factors, to ensure the scientificity of data and to avoid the bias caused by subjective arrangement. Therefore, we randomly divided the 50 samples of the study into two groups, one as the testing group and the other as the validating group. The 30-sample testing group was used to construct the prediction and alert model, while the 20-sample validating group was used to verify the accuracy of the algorithm and modify the model according to the analysis results.

An example of how the data are related to online and offline learning is the scenario dialogue activity. The students would assign roles after the teacher released the activity on the platform. They can either role play dialogue online or practice with group members offline before recording the dialogue on the platform. The platform keeps data of the students' voices, number of practice sessions, and scores. The data in Figures 2 and 3 are from the platform.

4. scene dialogue new standard for external research grade 8 Volume 1 modules 9 Average score 85. pass rate 92%

Average	Scripts
83 points	Did you finish the report called "Chinese population", Helen?
89 points	No, but I'm getting ready for it.
87 points	Beijing has a population of about twenty million. That causes lots of problems.
83 points	I'm afraid that it's not only Beijing. Do you know which city has the largest population in China?
87 points	Yes, it's Chongqing. About more than thirty-three million people live there.
84 points	I can't believe it!
84 points	The population of China is about 1.37 billion. That's almost one fifth of the world's population.
84 points	Wait for a moment! I will write that down! But how do you know all this?
85 points	I finished my report yesterday!
84 points	Great, thank you for telling me that!

Figure 2. The “imitate and utter the sentence” example.

100%	90	90	05:00:02–2010:52 no comment issued
100%	85	85	12 min 47 s 02–2010:52 no comment issued
100%	92	92	0707:43:02–2020:55 no comment issued
100%	94	94	07:38:02–2010:57 no comment issued
100%	82	82	060602–2010:59 no comment issued

Figure 3. The submission time of the practice.

In Figure 2, the right side is the specific content of the students’ exercises, and the left side is the score given by the teacher according to the completion degree of the students’ exercises.

The purpose of the activity is to practice English speaking. We judged the activity to happen in the classroom because the students submitted their practices nearly simultaneously during the morning class period. The data on the platform were categorized by the type of learning time and activity in Table 2.

Table 2. Data Classification on Hybrid Learning Platform.

Type of Activity	Activity	Time of Release and Submission	Type of Data
Listening	Listen and complete the test	Outside class	Online data
	Listen to the materials and choose the answer.	In and outside class	Classroom data and online data
Speaking	Read after the dialogue	In and outside class	Classroom data and online data
	Read after the text	In and outside class	Classroom data and online data
	Recite the text	Outside class	Online data
	Find and read the target language	In and outside class	Classroom data and online data
	Read and memorize the target sentence	In and outside class	Classroom data and online data
	Imitate and utter the sentence	In class	Classroom data
	Scenario dialogue	In class	Classroom data
Reading	Reading comprehension	Outside class	Online data
	Complete the dialogue	In and outside class	Classroom data and online data
Writing	Written expression	Outside class	Online data
	Complete the sentence	Outside class	Online data
	Combine paragraphs into a text	Outside class	Online data

To ensure the generalizability of the prediction and alert algorithms, 50 students were randomly selected from 11 eighth grade classes. Their hybrid learning data were then used for algorithmic analysis. The data to be analyzed had six attributes that were divided into two categories—academic performance and academic behavior—as shown in Table 3 [15].

Table 3. Data Attributes on Hybrid Learning Platform.

Category	Attribute	Attribute Description
Learning achievements	Qualified rate	Varies from 1% to 100%
	Excellent rate	Varies from 1% to 100%
	Scores	Varies from 1 to 100
	Completion	Varies from 1% to 100%
Learning behaviors	Number of practice sessions	Varies from 1min to + ∞ min
	Practice time	Varies from 1 to + ∞ times

The sample was selected from eighth-grade students, and learning data are generated throughout the students' learning process, so as long as the students continue to use the platform, the data generated are theoretically unlimited.

3.2. Determination of Alert Attributes

As demographic and education background attributes are fixed values, and the learning performance is largely represented by the student's "scores", the correlation analysis was thus conducted between the variable of "scores" and the two categories of attribute data, namely, "qualified rate", "excellent rate", "completion", "practice time", and "number of practice sessions". The early alert indicators are determined by the significance of the correlation coefficients.

As Table 4 shows, the Pearson correlation coefficients between "scores" and "pass rate", "excellence rate", "practice time", "completion", and "number of practice sessions" are 0.935, 0.846, 0.849, 0.817, and 0.758, respectively, all of which are significant at the 0.01 level, indicating a significant positive correlation of "scores" to the other attributes. Therefore, "qualified rate", "excellent rate", "practice time", "completion", and "number of practice sessions" are determined to be the alert attributes of a hybrid learning performance.

Table 4. Correlations of the Indicators to "scores".

Pearson Correlation		
		Scores
Qualified rate	Correlation Coefficient	0.935
	<i>p</i> -value	0.000
Excellent rate	Correlation Coefficient	0.846
	<i>p</i> -value	0.000
Number of practice sessions	Correlation Coefficient	0.849
	<i>p</i> -value	0.000
Completion	Correlation Coefficient	0.817
	<i>p</i> -value	0.000
Practice time	Correlation Coefficient	0.758
	<i>p</i> -value	0.000

3.3. Construction of Ridge Regression Prediction Model

Ridge regression is a statistical method to solve multicollinearity among independent attributes, which eliminates data multicollinearity to establish a stable model [24]. To be more specific, it takes care of ill-conditioned data by getting rid of some weak relative data to ensure the accurate estimates of the regression coefficients. Therefore, it does create significant *p*-values and enhance the predictability of the model.

In the linear regression model, the objective function is:

$$U(\alpha) = \sum (y - X\alpha)^2 + k\|\alpha\|_2^2$$

while in the ridge regression model, a penalty term of L2 norm is added to the objective function to attain the regression coefficients α , as follows:

$$\begin{aligned} U(\alpha) &= \sum (y - X\alpha)^2 \\ &= \sum (y - X\alpha)^2 + \sum k\alpha^2 \end{aligned}$$

where k is a non-negative value. The larger k is, the smaller $U(\alpha)$ and the coefficient α are. The key to ridge regression is to balance the model's biases with a reasonable k value.

In the research, we imported the data of the 30-sample testing group into Python and converted the data into a matrix for calculation, some of which are illustrated in Table 5 and Figure 4.

Table 5. Partial tabular data of the testing group.

Flat Number	Rate Qualified	Rate Excellent	Rate Practice	Times	Rate Completion	Time
1	50.07	60	9	5	25	0.38
2	45.76	66	2	5	25	0.46
3	35.64	56	43	13	87	5
4	80.09	87	53	15	94	2.3
5	65.65	64	29	14	88	3
6	40.81	33	0	3	15	0.16

1	52	87	60	0	5	25	0	38
2	45	76	60	0	5	25	0	46
3	83	59	85	62	13	87	0	5
4	80	9	87	53	15	94	2	3
5	65	65	64	29	14	88	0	3
6	40	81	33	0	3	15	0	16

Figure 4. Partial matrix data of the testing group.

We started modeling in Python in the following steps: apply the ridge regression algorithm in Python and encapsulate the least-squares method; observe the curve slope of each independent attribute as the value of K is taken successively from the smaller to the larger; watch for the collinearity and the best value of K to draw the ridge trace diagram, as shown in Figure 5.

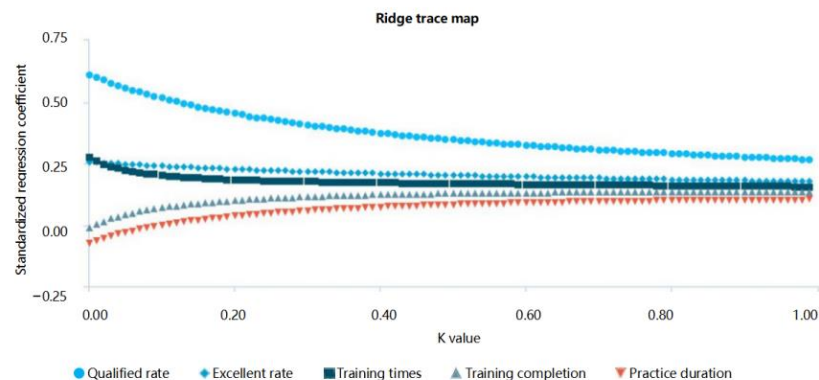


Figure 5. Schematic diagram of ridge trace map.

With the increase in the K value, the curve slope of each independent attribute gradually levels off. When K is 0.99, the slope of each attribute is the most stable, indicating $K = 0.99$ as the optimal value. After the optimal value of K is determined, the 30 matrix data of the testing group are imported into Python to construct a ridge regression prediction model, outputting the R^2 -value, intercept, and the coefficient matrix of each attribute, as shown in Figure 6.

```
Intercept:42.148661
Coefficient matrix:
[0.19029897 0.08485868 0.44081529 0.0614919 1.028915]
Using regression model:
Ridge(alpha=0.99, copy_X=True, fit_intercept=True, max_iter=None,
normalize=True, random_state=None, solver='auto', tol=0.001)
Training set test effect R2:
0.86623763536983
```

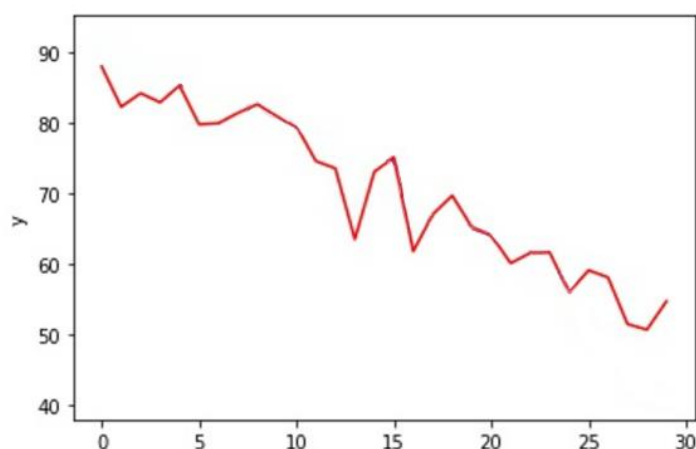


Figure 6. Ridge regression algorithm.

As Figure 6 shows, when $K = 0.99$, the model reaches the maximum fitting. The attained formula for learning performance prediction is shown as follows:

$$S_z = 42.149 + 0.19L_h + 0.085L_y + 0.441E_n + 0.061E_f + 1.029E_t$$

In order to explore whether each independent attribute has a significant positive influence on the dependent attribute, the study conducted an F -test on the ridge regression prediction model, the results of which are shown in Table 6.

Table 6. F -Test on the Ridge Regression Prediction Model.

	Normalization Factor β	T -Value	p -Value	R^2	F
Constant	-	17.017	0.000	0.866	$F(5,24) = 31.084$, $p = 0.000$
Qualified rate	0.266	8.259	0.000		
Excellent rate	0.176	5.873	0.000		
Number of practice sessions	0.155	6.194	0.000		
Completion	0.133	5.000	0.000		
Practice time	0.102	3.553	0.002		

According to Table 5, all independent attributes show a significant positive influence on the dependent attribute “scores”.

3.4. Construction of K-Means Clustering Alert Model

Means clustering is used to solve classification problems. As a distance-based clustering algorithm, it follows the principle that the shorter the distance between two objects, the higher the similarity between them. K-means clustering aims to obtain independent clusters, each of which consists of objects close to each other. In the application, the convergence is fast, and the classification effect is obvious and better. The desired clusters can be obtained by simply adjusting the K-values of the centroids. The steps of the K-means clustering algorithm are first to randomly select K objects as the initial cluster centers, then to calculate the distance between each object and each seed cluster center, and finally, to assign each object to the nearest cluster center. One cluster center, together with the objects assigned to it, makes up a cluster. Whenever a sample is added, the cluster centers are recalculated based on the existing objects in the clusters. In each course, teachers’ requirements for students’ status are different, and, at the same time, teachers’ determination of students’ performance is also different; therefore, it is not reasonable to set a fixed alert line in the learning alert link. The k-means clustering algorithm can divide the sample data into corresponding clusters according to the researcher’s needs.

In order to predict learning performance, the study divides learning performance into three states: no alert, mild alert, and severe alert. Therefore, when clustering the learning performance of the testing group, the value of K was set as 3, and the sub-groups of no alert, mild alert, and severe alert were defined according to the central value of each cluster. Shown in Figure 7 are the results from the K-means clustering of the dependent attribute “scores” based on the data of the testing group.



```
Python 3.6.4 (v3.6.4:d48e3eb, Dec 19 2017, 06:54:40) [MSC v.1900 64 bit (AMD64)]
on win32
Type "copyright", "credits" or "license()" for more information.
>>
===== FSTART: C:\Users\heyde\Desktop\1111.py =====
[[41.42503455]
 [84.41667169]
 [69.31199247]]
[[45.01]
 [84.40666667]
 [66.81846154]]
```

Figure 7. K-means clustering based on the data of the testing group.

The sample data are divided into three categories after two iterations. The cluster centers of each category are 45.01, 84.4066667, and 66.81146154, respectively. The first category with five samples is defined as the severe alert group, the second category with 13 samples, the mild alert group, and the third category with 12 samples, the no alert group.

Of the 12 samples in the third category, the sample value farthest from the cluster center and less than the cluster center value 84.41 is 78.43. Therefore, the no alert line L1 is set at 78.43. The samples with higher scores than L1 are in the no alert zone. Of the five samples in the first category, the sample value farthest from the cluster center and less than the cluster center value 45.01 is 40.63. Therefore, the alert warning line L3 is set at 40.6. The samples whose scores are lower than L3 are all in the severe alert zone. Of the 13 samples in the second category, the sample value farthest from the cluster center and less than the cluster center value 66.82 is 57.84. Therefore, the mild alert line L2 is set at 57.84. The samples whose scores are higher than L3 and lower than L2 are all in the mild alert zone in Figure 8.

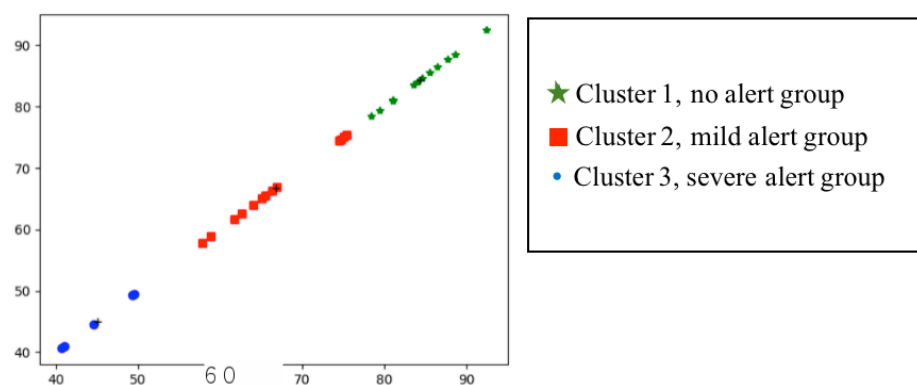


Figure 8. Clustering diagram.

3.5. Validation of the Overall Alert Accuracy

To verify the accuracy of the alert model, the 20-sample data in the validating group were brought to the prediction formula of the hybrid learning performance. Then, K-means clustering analysis was exerted on the two groups of data, with the K-value set as 3. The corresponding comparison between the two groups of data is illustrated in Figure 9, according to which, as the value along the ordinate axis decreases, the downward trends of both groups basically fit each other, except for s18, s19, and s20 in the severe alert zone. In terms of the number of members in each zone, the membership distribution in each zone is basically consistent except for a slight deviation in the severe alert zone, indicating a fitting effect of the alert model, as expected.

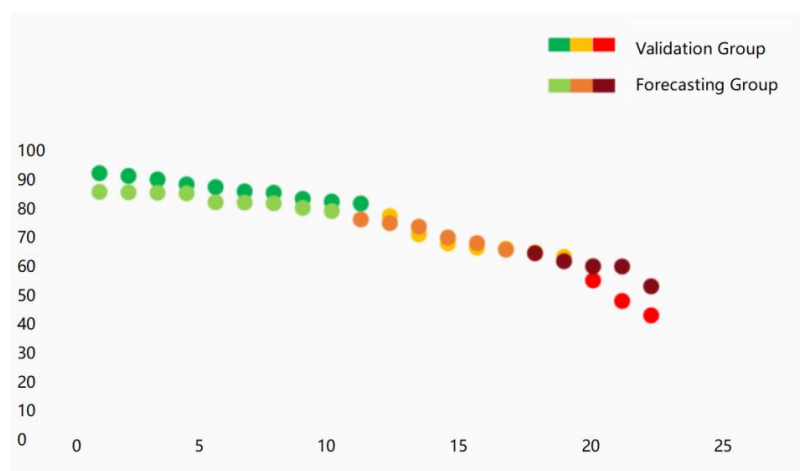


Figure 9. Accuracy verification diagram.

With v representing the number of samples that the model warns correctly, and m , the number of alert errors, the accuracy calculation formula is:

$$Accuracy = \frac{v}{v + m} = \frac{17}{17 + 3} = 0.85$$

Normally, the closer the accuracy value is to 1, the better the modeling effect is. The generated model proves the high quality, with an 85% accuracy according to the calculation.

To probe into the accuracy of the alert model at different time points, the study selected 20 samples of data from September to December 2019, respectively, from the testing and the validating groups, and compared the K-means clustering by month between the two groups, with the initial K-value set as 3, and the number of iterations, within 10. The results of the K-means clustering from September to December are shown in the following Figure 10 and Table 7.

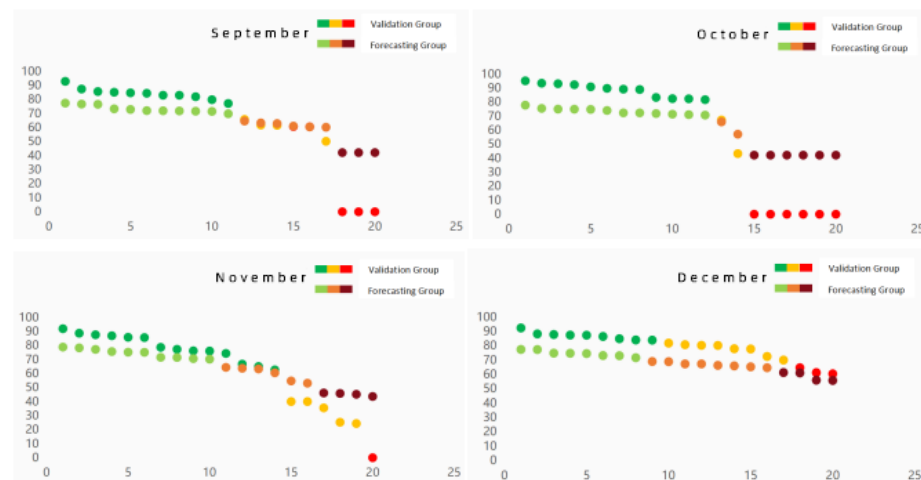


Figure 10. The charts of k-means clustering from September to December.

Table 7. K-Means Clustering from September to December.

Month	Testing Group	Number of Samples	Alert Line	Validating Group	Number of Samples	Alert Line	Accuracy
September	No alert zone	11	69.812	No alert zone	11	77.13	100%
	Mild alert zone	6	60.191	Mild alert zone	6	50.19	
	Severe alert zone	3	42.149	Severe alert zone	3	0.00	
October	No alert zone	12	70.728	No alert zone	12	81.75	100%
	Mild alert zone	2	57.102	Mild alert zone	2	43.25	
	Severe alert zone	6	42.149	Severe alert zone	6	0.00	
November	No alert zone	10	70.217	No alert zone	14	62.58	65%
	Mild alert zone	6	53.08	Mild alert zone	5	24.34	
	Severe alert zone	4	43.637	Severe alert zone	1	0	
December	No alert zone	8	71.609	No alert zone	9	83.86	85%
	Mild alert zone	8	64.64	Mild alert zone	8	69.96	
	Severe alert zone	4	55.735	Severe alert zone	3	60.39	

The results shown in Table 6 suggest a high accuracy of the alert model in most months. Except for November at 65%, the accuracy in the other three months exceeds 80%, with the accuracy in September and October as high as 100%. The accuracy of the alert model is thus basically up to expectation.

4. Discussion

In the alert process, there is a discrepancy between the alert lines in the validating group and those of the alert group nodes, especially in the severe alert zone. Figure 11 presents the chart comparing the data between the two groups (the ridge regression test set image generated in Python).

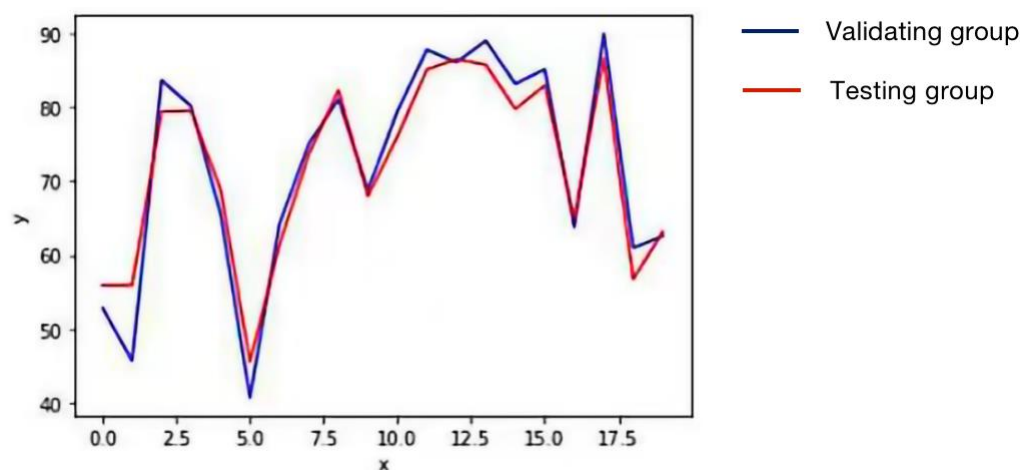


Figure 11. Data comparison between the validating group and the testing group.

In Figure 11, the blue curve represents the data of the validating group, while the red curve, the data of the testing group. According to the track coincidence degree of the red and blue curves, the data in the validating group are basically consistent with those in the testing group. However, in the high-score segment, the data of the validating group are slightly higher than those of the testing group, while in the low-score segment, they are slightly lower than those of the testing group. In the medium-score segment, the curves of the two groups almost completely fit each other.

Since the key alert objects are the learners in the severe alert zone, the data of both groups in low-score segments are further analyzed. It should be noted that there are zero-scored samples in the four-month data of the validating group, except for December. This means that the “qualified rate”, “excellent rate”, “number of practice sessions”, “completion”, and “practice time” of these samples are all 0. Their values in the testing group are equally set as 42.149, which is the intercept value (constant term) in the alert model. That is why the cluster center value of the validating group in the severe alert zone is much smaller than that of the testing group. Given this situation, the alert model is modified by adding a condition item, if the independent attributes of “qualified rate”, “excellent rate”, “number of practice sessions”, “completion”, and “practice time” are all 0, the learning performance value of the dependent attribute is also set as 0. The modified learning performance prediction model in hybrid learning is expressed as follows:

$$S_z = 42.149 + 0.19L_h + 0.085L_y + 0.441E_n + 0.061E_f + 1.029E_t$$

When L_h, L_y, E_n, E_f, E_t are equal to 0, $S_z = 0$

In order to verify the effectiveness of the adjusted alert model, the alert results formed in the previous month, from September to December, are compared with the samples’ real learning performances of the month, as is shown in Table 8.

In the no alert zone, the alert accuracy of the model is 75%, 80%, and 76.9% in October, November, and December, respectively, while in the alert zones, it is 75%, 62.5%, and 80%, respectively, suggesting an effectiveness of the model as expected. However, there is a certain confusion between the mild alert and severe alert zones. In combination with the actual teaching situations, the study used the data from the learning platform that reflect the students’ hybrid learning performance. They may represent the reduced learning performance of some students not only on the platform but also in the classroom, which could be noticed by the teacher and then coped with accordingly. The students may also adjust their learning performance under the teacher’s assistance and guidance. Therefore, the actual learning effects differ from the prediction, which is also the significance of learning performance prediction and alert per se. However, the teacher’s guidance or the student’s adjustment may vary from person to person with different learning outcomes.

This may also lead to a deviation in learning performance predication and the confusion of mild and severe alert zones in the data analysis.

Table 8. Comparison of the Alert and the Actual Situations.

Alert situation			Actual situation		
September	No alert	s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11	October	No alert	s1, s2, s3, s4, s6, s7, s8, s9, s10, s11, s14, s17
	Mild alert	s12, s13, s14, s15, s16, s17		Mild alert	s5, s12
	Severe alert	s18, s19, s20		Severe alert	s13, s15, s16, s18, s19, s20
October	No alert	s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11, s12	November	No alert	s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11, s12, s13, s14, s18
	Mild alert	s13, s14		Mild alert	s15, s16, s19, s20
	Severe alert	s15, s16, s17, s18, s19, s20		Severe alert	s17
November	No alert	s1, s2, s3, s4, s5, s6, s7, s8, s9, s10	December	No alert	s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11, s12, s18
	Mild alert	s11, s12, s13, s14, s15, s16		Mild alert	s13, s14, s15, s16, s17, s19, s20
	Severe alert	s17, s18, s19, s20		Severe alert	s12

“s” = “student”.

5. Conclusions

After the outbreak of COVID-19, hybrid learning has become the mainstream learning method, and the quality of hybrid learning has become particularly important. In this research, a “prediction-alert” model based on a ridge regression algorithm and a K-means clustering algorithm can effectively predict students’ learning performance by integrating online and offline learning data. A valid learning performance prediction and alert model contributes to the effectiveness and the accuracy of teaching in a hybrid learning context. It facilitates teachers’ analysis of the students’ learning performance, the choice of teaching strategies, and the implementation of precise teaching. It also proves a feasible method for the analysis of hybrid learning performance.

With the cross-fertilization of 5G, artificial intelligence, big data, the Internet of Things, and other new information technologies with education at multiple levels and angles, education development has also accelerated from digitalization and networking to intelligence. Strengthening the application of big data and artificial intelligence technology in the field of learning monitoring and early alerts can promote educational research to advance deeply in the direction of digitalization and intelligent innovation. In addition, it can continuously improve the accuracy and personalization of education services and realize a data-driven learner-centered teaching model. It also helps educational decision making move toward digitization, intelligence, and precision. This study explores how to integrate log data, behavior data, and operating data to provide early alerts of learning performance in a hybrid learning environment, and how to design an early alert system for dynamic monitoring of the learning process to provide an effective basis for precision teaching and learning interventions.

Author Contributions: Data curation, J.D.; Project administration, H.L.; Resources, H.Z.; Writing—review & editing, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: 2020 Youth Fund of Ministry of Education, Humanities and Social Science Research Project “Research on the Local Internationalization Talents Cultivation based on the Global Competency” (20YJC880135).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. U.S. Department of Education Office of Educational Technology. The National Education Technology Plan. 2017. Available online: <https://tech.ed.gov/netp/> (accessed on 15 January 2017).
2. Chen, X.; Xie, H.; Zou, D.; Hwang, G.J. Application and theory gaps during the rise of artificial intelligence in education. *Comput. Educ. Artif. Intell.* **2020**, *1*, 100002. [CrossRef]
3. Zhai, X.; Chu, X.; Chai, C.S.; Jong, M.S.Y.; Istenic, A.; Spector, M.; Liu, J.B.; Yuan, J.; Li, Y. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* **2021**, *2021*, 8812542. [CrossRef]
4. Rasheed, R.A.; Kamsin, A.; Abdullah, N.A. Challenges in the online component of blended learning: A systematic review. *Comput. Sci.* **2020**, *144*, 103701. [CrossRef]
5. Klimova, B.F.; Kacetl, J. Hybrid learning and its current role in the teaching of foreign languages. In Proceedings of the 4th World Conference on Educational Technology Researches (WCETR 2014), Barcelona, Spain, 13 May 2015; pp. 477–481. [CrossRef]
6. Navio-Marco, J.; Ruiz-Gomez, L.M.; Arguedas-Sanz, R.; Lopez-Martin, C. The student as a prosumer of educational audio-visual resources: A higher education hybrid learning experience. *Interact. Learn. Environ.* **2022**, *Latest articles*, 1–18. [CrossRef]
7. Liu, J.L.; Yang, Z.T.; Wang, X.; Zhang, X.R.; Feng, J.Y. An Early-Warning Method on e-Learning. In Proceedings of the 4th EAI International Conference on e-Learning, e-Education, and Online Training (eLEOT 2018), Shanghai, China, 5–7 April 2018; pp. 62–72. [CrossRef]
8. Hwang, G.J.; Xie, H.; Wah, B.W.; Gašević, D. Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Comput. Educ. Artif. Intell.* **2020**, *1*, 100001. [CrossRef]
9. Romero, C.; Ventura, S. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Sys. Man Cyber. Part C-App. Rev.* **2010**, *40*, 601–618. [CrossRef]
10. Félix, C.; Alfredo, V.; Angela, N.; Francisco, M. Applying Data Mining Techniques to e-Learning Problems. In *Evolution of Teaching and Learning Paradigms in Intelligent Environment*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 183–221. [CrossRef]
11. Gong, S.; Qin, X. Research on Real-time Learning Prediction Method Based on Spark. In Proceedings of the 10th International Conference on Software Engineering and Service Science (ICSESS 2019), Beijing, China, 18–20 October 2019; pp. 354–357. [CrossRef]
12. Wei, N. A Data Mining Method for Students’ Behavior Understanding. *Int. J. Emer. Techno. Learn.* **2020**, *15*, 18–32.
13. Lu, O.; Huang, A.; Huang, J.; Lin, A.; Ogata, H.; Yang, S. Applying Learning Analytics for the Early Prediction of Students’ Academic Performance in Blended Learning. *Educ. Technol. Soc.* **2018**, *21*, 220–232. Available online: <http://www.jstor.org/stable/26388400> (accessed on 30 April 2018).
14. Riestra-González, M.; del Puerto Paule-Ruiz, M.; Ortin, F. Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Comput. Educ.* **2021**, *163*, 104108. [CrossRef]
15. Zhai, J.H.; Zhu, Z.Z.; Li, D.Q.; Huang, N.X.; Zhang, K.Y.; Huang, Y.Q. A Learning Early-Warning Model Based on Knowledge Points. In Proceedings of the 15th Conference of Intelligent Tutoring Systems (ITS 2019). Suntec, Singapore, 21 October 2019; pp. 1–6. [CrossRef]
16. Hu, Y.H.; Lo, C.L.; Shih, S.P. Developing early warning systems to predict students’ online learning performance. *Comput. Hum. Behav.* **2014**, *36*, 469–478. [CrossRef]
17. Akcapinar, G.; Hasnine, M.N.; Majumdar, R.; Flanagan, B.; Ogata, H. Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learn. Environ.* **2019**, *6*, 4–15. Available online: <https://slejournal.springeropen.com/articles/10.1186/s40561-019-0083-4> (accessed on 10 May 2019). [CrossRef]
18. Howard, E.; Meehan, M.; Parnell, A. Contrasting prediction methods for early warning systems at undergraduate level. *Internet High. Educ.* **2018**, *37*, 66–75. [CrossRef]
19. Cheng, X.X.; Zhu, Z.Z.; Liu, X.; Yuan, X.F.; Guo, J.Y.; Guo, Q.; Li, D.Q.; Zhu, R.F. A Novel Learning Early-Warning Model Based on Random Forest Algorithm. In Proceedings of the 14th International Conference on Intelligent Tutoring Systems (ITS 2018), Copenhagen, Denmark, 17–21 September 2018; pp. 306–312. [CrossRef]
20. Jiang, Z.; Zhang, Y.; Li, X. Learning Behavior Analysis and Prediction Based on MOOC Data. *J. Comput. Res. Dev.* **2015**, *52*, 614–628. [CrossRef]
21. Limsathitwong, K.; Tiwatthanont, K.; Yatsungnoen, T. Dropout Prediction System to Reduce Discontinue Study Rate of Information Technology Students. In Proceedings of the 5th International Conference on Business and Industrial Research (ICBIR 2018), Bangkok, Thailand, 17–18 May 2018; pp. 110–114. [CrossRef]

-
22. Xu, X.; Wang, J.; Peng, H.; Wu, R. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Comput. Hum. Behav.* **2019**, *98*, 166–173. [[CrossRef](#)]
 23. Matzavela, V.; Alepis, E. Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100035. [[CrossRef](#)]
 24. Dorugade, A.V. New Ridge Parameters for Ridge Regression. *J. Asso. Arab Univ. Basic Appl. Sci.* **2014**, *15*, 94–99. [[CrossRef](#)]